

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-259138

(43)Date of publication of application : 03.10.1997

(51)Int.Cl.

G06F 17/30

G06F 12/00

(21)Application number : 08-065000

(71)Applicant : N T T DATA TSUSHIN KK

(22)Date of filing : 21.03.1996

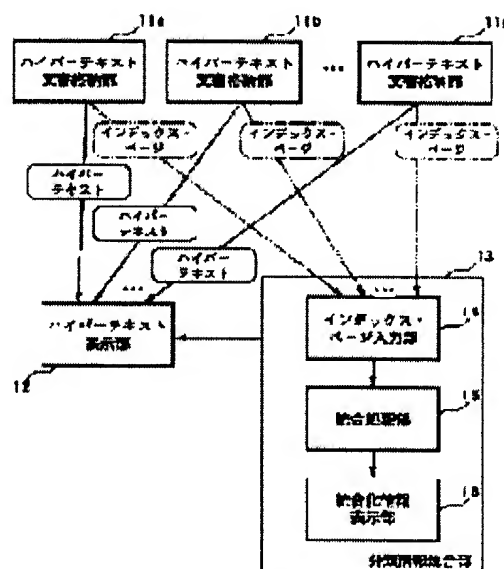
(72)Inventor : NONAKA SATORU

(54) SORT INFORMATION DISPLAY METHOD AND INFORMATION RETRIEVAL DEVICE

(57)Abstract:

PROBLEM TO BE SOLVED: To improve the retrieval efficiency of hypertext information that is sorted and systematized from different points of view and also to improve the operability in a retrieval mode.

SOLUTION: Plural index pages (sort information) acquired from hypertext document storage parts 11a, 11b...11n are inputted to a sort information integration part 13. The part 13 extracts the category names (sort items), link names and link destination document IDs (URLs) out of those index pages and integrates them. At the same time, the part 13 statistically analyzes its integrated information based on every category name and document ID and makes clear the similarity relations among categories and link destination documents. Then, the part 13 shows the similar categories and link names on the screen of a hypertext display part 12 in an integrated way based on the degrees of similarity.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平9-259138

(43)公開日 平成9年(1997)10月3日

(51)Int.Cl. ⁹	識別記号	序内整理番号	F I	技術表示箇所
G 0 6 F	17/30		G 0 6 F	15/403
	12/00	5 4 7		12/00
				15/40
				15/403
				15/419
				3 7 0 Z
				5 4 7 H
				3 1 0 C
				3 8 0 D
				3 2 0
審査請求 未請求 請求項の数 6 O L (全 7 頁)				

(21)出願番号 特願平8-65000

(22)出願日 平成8年(1996)3月21日

(71)出願人 000102728

エヌ・ティ・ティ・データ通信株式会社
東京都江東区豊洲三丁目3番3号

(72)発明者 野中 哲

東京都江東区豊洲三丁目3番3号 エヌ・
ティ・ティ・データ通信株式会社内

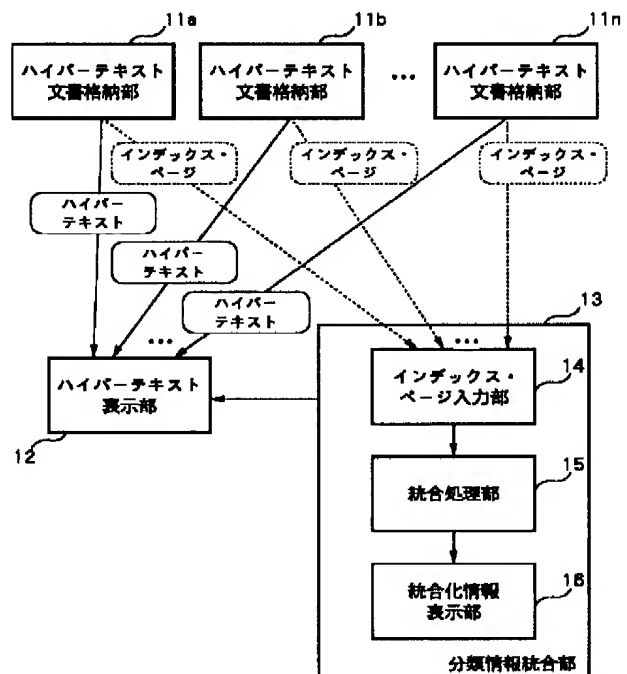
(74)代理人 弁理士 鈴木 正剛

(54)【発明の名称】 分類情報表示方法及び情報検索装置

(57)【要約】

【課題】 異なる観点で分類／体系化されたハイパーテキスト情報の検索効率を高めるとともに、検索時の操作性をも向上させる情報検索装置を提供する。

【解決手段】 ハイパーテキスト文書格納部11A, 11B, …11nから取得した複数のインデックス・ページ(分類情報)を分類情報統合部13に入力する。分類情報統合部13は、各インデックス・ページに含まれるカテゴリ名(分類項目)、リンク名、リンク先文書ID(URL)を抽出して統合するとともに、各カテゴリ名とリンク先文書IDをもとに、統合した情報を統計的に分析して個々のカテゴリ間やリンク先文書間の類似関係を明らかにする。そして、類似するカテゴリやリンク名を、類似度の尺度に応じてハイパーテキスト表示部12の画面上に統合表示する。



【特許請求の範囲】

【請求項1】 検索対象データの分類を表すカテゴリー情報と個々の検索対象データの格納先を指標するリンク情報とを含む複数の分類情報を統合し、統合した分類情報に含まれる個々のカテゴリー情報とリンク情報とを各々の相関係数値に応じた間隔で同一画面に統合表示することを特徴とする分類情報表示方法。

【請求項2】 検索対象データの分類を表すカテゴリー情報と個々の検索対象データの格納先を指標するリンク情報とを含む複数の分類情報を分析する情報検索装置における分類情報表示方法であって、前記情報検索装置が、

取得した複数の分類情報の各々に含まれるカテゴリー情報及びリンク情報を抽出する第1ステップと、抽出した情報間の類似度を統計的に導出する第2ステップと、前記抽出した個々の情報をそれぞれ前記導出した類似度に応じた尺度で統合表示する第3ステップと、を少なくともこの順に実行することを特徴とする検索情報表示方法。

【請求項3】 前記第2ステップが、個体数量及び特性数量をパラメータに含む「数量化III類」を用い、前記リンク情報を前記個体数量、前記カテゴリー情報を前記特性数量にそれぞれ対応させる過程を含むことを特徴とする請求項2記載の分類情報表示方法。

【請求項4】 前記個体数量及び特性数量に基づく統計結果をそれぞれ前記リンク情報及びカテゴリー情報に対応するベクトル情報として定量化する過程を含むことを特徴とする請求項3記載の分類情報表示方法。

【請求項5】 前記検索対象データがハイパーテキスト情報であり、前記リンク情報が前記ハイパーテキスト情報のリンク先IDを含むことを特徴とする請求項1ないし4のいずれかの項記載の分類情報表示方法。

【請求項6】 分散配置されている複数の情報格納部から検索対象データの分類を表すカテゴリー情報及び個々の検索対象データの格納先を指標するリンク情報を含む複数の分類情報を取得する分類情報取得手段と、この分類情報取得手段で取得した複数の分類情報を統合するとともに、各分類情報に含まれるカテゴリー情報及びリンク情報を、それぞれ統計的に導出した各情報間の類似度に応じて定量化する統合処理手段と、前記定量化されたカテゴリー情報及びリンク情報をそれぞれ前記類似度に応じた尺度で統合表示する表示手段と、を有する情報検索装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、例えばインターネット上のハイパーテキスト情報検索システムであるWWW(World Wide Web)のような複数の情報源の検索技術に係り、特に、異なる情報提供者が各々

独自の観点で分類した情報が複数存在する場合の分類情報の分析技術に関する。

【0002】

【従来の技術】今日、WWWの普及により、無数の情報源から多種多様な情報を得ることが可能になっている。WWWは、図4に示すように、WWWサーバに相当する複数のハイパーテキスト文書格納部HPM1、HPM2…HPMnと、WWWブラウザに相当するハイパーテキスト表示部HPDとを備えて構成される。ハイパーテキスト文書格納部HPM1、HPM2…HPMnには、検索対象データとなる複数のハイパーテキスト文書が格納されている。各ハイパーテキスト文書は、ハイパーテキスト形式、即ちHTMLと呼ばれるマークアップ言語で記述され、リンク先の文書IDであるURLにより一意にその格納先が特定できるようになっている。

【0003】ハイパーテキスト文書格納部HPM1、HPM2…HPMnには、また、情報提供者が予めURLを用いて分類するとともに、カテゴリー名、リンク名、ハイパーリンク等を付与して、利用者（情報検索者、以下同じ）が所望のハイパーテキスト文書を検索をしやすい形にした各種分野のインデックス・ページ（分類情報）が存在する。インデックス・ページの代表例としては、周知の「Yahoo」が挙げられる。このインデックス・ページもまた、HTMLにより記述されるものである。

【0004】利用者は、まず、通信手段を用いて上記インデックス・ページを取得し、これをハイパーテキスト表示部HPDを通して表示して、どのインデックス・ページの、どのカテゴリーに、どのようなハイパーテキスト文書がリンクしているかを知る。図5は、インデックス・ページに基づくハイパーテキスト表示部HPDの表示例を示す図であり、図6はこのときのHTMLによる記述例を示す図である。

【0005】図5において、「〇〇技術」や「××技術」がカテゴリー名、「A会社」や「B研究所」等がリンク名、個々のリンク名に付された下線がハイパーリンクである。図示の例では、「〇〇技術」のカテゴリーに「A会社」、「B研究所」に関する文書がリンクし、「××技術」のカテゴリーに「C製品」、「D研究所」に関する文書がリンクしていることを示している。また、利用者が、ハイパーリンクが付されたテキストあるいは画像等をポインティング・デバイス等で選択することにより、対応するハイパーテキスト文書にアクセスできるようになっている。

【0006】

【発明が解決しようとする課題】ところで、WWWの専門的な分野のインデックス・ページの中には、異なる情報提供者が各々独自の観点で分類／体系化して整理したものがあ

ことは必ずしも容易でない。そのため、複数のインデックス・ページとURLやリンク名等のリンク情報との関係が矛盾したり、リンク情報間で過不足が生じたりする場合がある。

【0007】図7は、この様子を示す説明図である。図中、符号(1)は、インデックス・ページAに存する「Aカンパニー」とインデックス・ページBに存する「A会社」のように、URLであるリンク先の文書ID(001)が同一であってもリンク名が異なる様子を示しており、符号(2)は、「D研究所」という同一リンク名であってもリンク先の文書ID(002, 003)が異なる様子を示している。また、符号(3)は、同じ「〇〇技術」でも、インデックス・ページA, B間で、リンク名が完全に一致していない、即ちリンク名の過不足が生じていることを示している。

【0008】このような場合、利用者は、ハイパーテキスト表示部HPDで複数のインデックス・ページを繰り返し開いてリンク名等の関係を参照しなければならず、検索効率及び検索時の操作性が極めて悪いという問題があった。このような問題は、WWWのインデックス・ページを用いた情報検索に限らず、異なる情報提供者が独自の観点で分類／体系化された分類項目を有する他の種類のハイパーテキスト情報の検索システム、あるいはインターネットのURLと同様のリンク先の情報ID(ポインタ等)を用いて情報を検索するデータベースシステムに共通に存し、改善が望まれていた。

【0009】本発明の課題は、上記事情に鑑み、異なる観点で分類／体系化された複数の分類項目を有する情報の検索を効率的にするとともに、検索時の操作性をも向上させる技術を提供することにある。

【0010】

【課題を解決するための手段】異なる情報提供者が各々独自の観点で分類／体系化した分類項目を扱う場合は、例えばURLやリンク名等のリンク情報により指標される検索対象データがどのカテゴリーに属するかという情報だけでなく、近いか遠いかというカテゴリー間の類似関係や、複数の分類項目での各カテゴリーとリンク情報とが各々どのような関係にあるかを十分に分析・把握することが重要となる。

【0011】一般に、分類項目LKがカテゴリーXjより構成されている場合、Yiなるリンク情報を含む場合を"1"、含まない場合を"0"とした行列成分 δ_{ij} を用いることにより、全ての分類項目のカテゴリーとリンク情報との関係を表現することが可能である。いま、実際にリンク情報を有する階層構造のカテゴリーのみを扱い、しかも下位のカテゴリーのみを含んでいるものは除外するものとする、カテゴリー(X)とリンク情報(Y)との相関係数 ρ_{XY} は、上記行列成分 δ_{ij} を用いて表すことができる。即ち相関係数 ρ_{XY} が最大になるようにすれば、カテゴリー(X)とリンク情報(Y)とをそ

れぞれ対応付けながら分類することができる。このような手法は、「数量化III類」と呼ばれ、外的基準の無いパタン分類を行う統計手法として知られている。なお、「数量化III類」のアルゴリズムは、例えば「数量化理論と方法」(林 知己夫著、朝倉書店)及び「数量化とデータ処理」(駒澤 勉著、朝倉書店)等を参考にすることができる。本発明は、この「数量化III類」を応用した分類情報表示方法及び情報検索装置を提供するものである。

10 【0012】即ち、本発明が提供する第1の分類情報表示方法は、検索対象データの分類を表すカテゴリー情報と個々の検索対象データの格納先を指標するリンク情報とを含む複数の分類情報を統合し、統合した分類情報に含まれる個々のカテゴリー情報とリンク情報とを各々の相関係数値に応じた間隔で同一画面に統合表示することを特徴とする。

20 【0013】また、第2の分類情報表示方法は、検索対象データの分類を表すカテゴリー情報と個々の検索対象データの格納先を指標するリンク情報とを含む複数の分類情報を分析する情報検索装置における分類情報表示方法であって、前記情報検索装置が、取得した複数の分類情報の各々に含まれるカテゴリー情報及びリンク情報を抽出する第1ステップと、抽出した情報間の類似度を統計的に導出する第2ステップと、前記抽出した個々の情報をそれぞれ前記導出した類似度に応じた尺度で統合表示する第3ステップとを少なくともこの順に実行することを特徴とする。検索対象データは、例えばハイパーテキスト情報であり、前記リンク情報は前記ハイパーテキスト情報のリンク先IDを含むものである。

30 【0014】第2の分類情報表示方法において、第2ステップは、個体数量及び特性数量をパラメータに含む上記「数量化III類」を用い、前記リンク情報を前記個体数量、前記カテゴリー情報を前記特性数量にそれぞれ対応させる過程を含む。その際、前記個体数量及び特性数量に基づく統計結果をそれぞれ前記リンク情報及びカテゴリー情報に対応するベクトル情報として定量化し、画面表示等によって可視化できるようにする。

40 【0015】また、本発明が提供する情報検索装置は、分散配置されている複数の情報格納部から検索対象データの分類を表すカテゴリー情報及び個々の検索対象データの格納先を指標するリンク情報を含む複数の分類情報を取得する分類情報取得手段、例えば通信制御装置と、この分類情報取得手段で取得した複数の分類情報を統合するとともに、各分類情報に含まれるカテゴリー情報及びリンク情報を、それぞれ統計的に導出した各情報間の類似度に応じて定量化する統合処理手段と、前記定量化されたカテゴリー情報及びリンク情報をそれぞれ前記類似度に応じた尺度で統合表示する表示手段と、を備えて構成する。

50 【0016】

【発明の実施の形態】以下、本発明の実施の形態を図面に基づいて説明する。図1は、本発明をインターネット上のハイパーテキスト情報検索システムであるWWWに適用する場合の一形態を示す構成図である。

【0017】図1において、符号11a, 11b, ..., 11nはハイパーテキスト文書格納部、12はハイパーテキスト表示部であり、それぞれ図5に示したハイパーテキスト文書格納部HPM1, HPM2...HPMn、ハイパーテキスト表示部HPDと同機能のものである。ハイパーテキスト文書格納部11a, 11b, ..., 11nには、HTMLで記述された複数のハイパーテキスト文書（検索対象データ）と、各種分野のインデックス・ページ（分類情報）が格納されている。インデックス・ページは、複数の情報提供者がそれぞれ独自の観点でURLやリンク名（リンク情報）を分類するとともに、分類項目名であるカテゴリー名（カテゴリー情報）を付与して、利用者がハイパーテキスト文書を検索をしやすい形にしたものである。

【0018】符号13は、本発明の情報検索装置の要部をなす分類情報統合部である。この分類情報統合部13は、ハイパーテキスト文書格納部11a, 11b, ..., 11nから複数のインデックス・ページを取得するとともに、取得したインデックス・ページに含まれる複数のカテゴリー情報やリンク情報を統合表示して検索処理の効率を高めるようにしたものである。具体的には、命令構成体であるプログラムによって機能形成されたインデックス・ページ入力部14、統合処理部15、及び統合化情報表示部16をこの順に縦続して成る。分類情報統合部13は、ハイパーテキスト表示部12にも接続されており、各部14~16の実行過程ないし実行結果を適時表示できるようになっている。なお、分類情報統合部13の機能は、汎用コンピュータが読取可能な情報担体に静的に固定され、あるいは汎用コンピュータに接続された通信媒体を通じて動的に固定された命令群と、これらの命令群を事後的にメモリ上に展開して所要機能を形成する上記汎用コンピュータによっても実現が可能なものである。

【0019】インデックス・ページ入力部14は、通信回線を通じてハイパーテキスト文書格納部11a, 11b, ..., 11nから複数のインデックス・ページが入力されたときに、各インデックス・ページに含まれるカテゴリー名、リンク名、リンク先文書IDを抽出するものである。入力された全てのインデックス・ページに対してこの処理を自動的に行った後、抽出した情報を統合処理部15に送る。

【0020】統合処理部15は、上記抽出情報を統合するとともに、各カテゴリー名とリンク先文書IDをもとに、統合した情報を上述の「数量化III類」を用いて統計的に分析し、個々のカテゴリー間やリンク先文書間の類似関係をベクトル情報の形で定量化するものである。

【0021】統合化情報表示部16は、統合処理部15で得られた全てのカテゴリーやリンク先文書についてのベクトル情報を、2次元あるいは3次元グラフ等の統合化情報としてハイパーテキスト表示部12の画面上に表示するための制御を行う。この統合化情報表示部16では、インデックス・ページ入力部14で得たリンク名を文書のベクトル情報に付与し、ハイパーテキスト表示部12に表示されたリンク名をポインティング・デバイスでクリックした場合に、該当する情報を表示する機能も備えている（ハイパーリンク）。

【0022】次に、本実施形態による情報検索過程を、より詳細に説明する。利用者は、まず、HTTPプロトコルにしたがう通信装置を用いてハイパーテキスト文書格納部11a~11nから複数のインデックス・ページを取得する。取得された各インデックス・ページは、分類情報統合部13のインデックス・ページ入力部14に入力され、ここでカテゴリー名、リンク名、リンク先文書IDが抽出される。抽出された情報は、統合処理部15において、個々のカテゴリーjがそれぞれどの文書iを含んでいるかを表す下記数1式のデータ行列Dの形で統合され、統計的に分析される。

【0023】

【数1】

$$D = (d_{ij}) = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{np} \end{pmatrix}$$

$$d_{ij} \rightarrow \begin{cases} 1 & (i \text{ 文書が } j \text{ カテゴリーに含まれる}) \\ 0 & (i \text{ 文書が } j \text{ カテゴリーに含まれない}) \end{cases}$$

【0024】この統合処理部15の処理手順を図2を参照して説明する。図2において、Sは処理ステップを示す。統合処理部15では、まず、カテゴリー及びリンク先文書ID（図中、文書IDと略称、以下同じ）の相互関係を表すデータ行列Dを得る（S1）。このデータ行列Dは、上記数1式により表されるものである。次に、S1で得た文書IDを個体数量（サンプルスコアともいう）、カテゴリーを特性数量（変数スコアともいう）にそれぞれ対応させ、「数量化III類」を用いて、各数量の算出を行う（S2）。S2で得た特性数量及び個体数量を、分散を考慮して正規化した後（S3）、特性数量をカテゴリーのベクトル情報、個体数量を文書IDのベクトル情報にそれぞれ変換して統合化情報表示部16に出力する（S4）。ベクトル情報は、長さ、向きを表す定量化情報であり、他のベクトルとの距離が両者の類似度を表している。

【0025】統合化情報表示部16は、統合処理部15で得られた各ベクトル情報に基づいて個々のカテゴリー間、文書ID間の類似度を画面上の距離情報に変換し、その統合表示制御情報を生成してハイパーテキスト表示

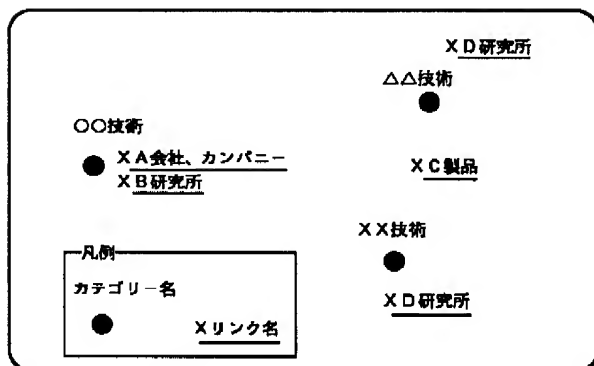
部12に送る。ハイパーテキスト表示部12は、統合表示制御情報に基づいて通常のインデックス・ページの場合と同様の手順で統合表示を行う。

【0026】ハイパーテキスト表示部12に統合表示される画面例を図3に示す。図3において、各カテゴリ間の距離、リンク名間の画面上の距離は、カテゴリ間の類似度、リンク先文書ID間の類似度、カテゴリとリンク先文書IDとの間の類似度をそのまま表しており、類似度が高い場合、あるいは同意義の場合は並記して表示される。例えば、「〇〇技術」のカテゴリに属する「A会社」と「Aカンパニー」は同意義であること、一方、「D研究所」はそれぞれ「△△技術」のカテゴリと「××技術」のカテゴリに同一リンク名で属しているが、両者は非類似であること、「C製品」は「△△技術」のカテゴリと「××技術」のカテゴリの双方に類似していること、「B研究所」は「〇〇技術」のカテゴリのみに属し、他のカテゴリとは関連しないこと、「△△技術」と「××技術」は比較的類似するが、「〇〇技術」との類似度が相対的に低いこと等をこの統合表示画面は示している。

【0027】このような統合表示画面を用いることにより、従来のようにリンク名である用語の整合性を考慮しながら複数のインデックス・ページを順次検索する操作を回避することができ、検索時の操作性を格段に向上させることができる。また、図3から明らかなように、用語間の類似関係が統合表示画面から直感的に判るため、利用者に当該分野に関する用語（概念）関係の理解を深める機会を与えることも可能になる。利用者が統合表示画面内の特定のリンク名（ハイパーリンク）をポインティングデバイス等で選択した場合は、該当文書が、ハイパーテキスト表示部12に表示される。

【0028】なお、以上は、カテゴリ情報及びリンク情報間の類似度を求めてその尺度に応じて統合表示する場合の処理について説明したが、「数量化III類」によ

【図3】



り個々のカテゴリ情報とリンク情報とを各々の相関係数値に応じた間隔で同一画面に統合表示する簡略化された手法を採用することもできる。また、図3では、2次元内における表現例を示したが、仮想空間技術を用いて3次元空間上で表現させるようにしてもよい。

【0029】

【発明の効果】以上の説明から明らかなように、本発明によれば、異なる情報提供者が独自の観点で分類した情報が複数存在する場合においても、各情報が統計的手法により分析され、それぞれの相関係数値あるいは類似度の尺度に応じて統合表示されるので、所望の情報の検索が従来に比べて簡略化されるとともに、検索時の操作性が格段に高まる効果がある。また、分類されている用語間の関係が直感的に明らかになるため、その分野における利用者の用語関係の知識獲得に貢献できるという利点も生じる。

【図面の簡単な説明】

【図1】本発明の実施の形態を示すブロック構成図。

【図2】統合処理部の内部処理の手順を示す図。

【図3】統合化された情報の表示例を示す説明図。

【図4】WWWの概念説明図。

【図5】インデックス・ページの例を示す説明図。

【図6】ハイパーテキスト形式であるHTMLの記述例を示す説明図。

【図7】従来例の問題点の説明図。

【符号の説明】

11A, 11B, ... 11n ハイパーテキスト文書格納部

12 ハイパーテキスト表示部

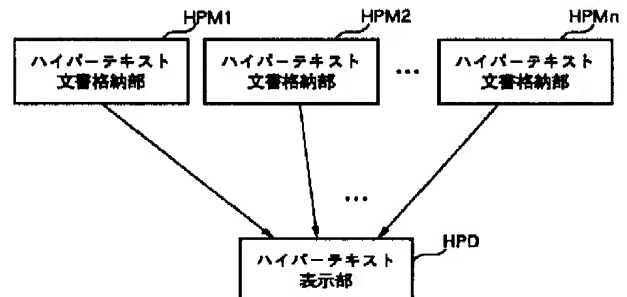
13 分類情報統合部

14 インデックス・ページ入力部

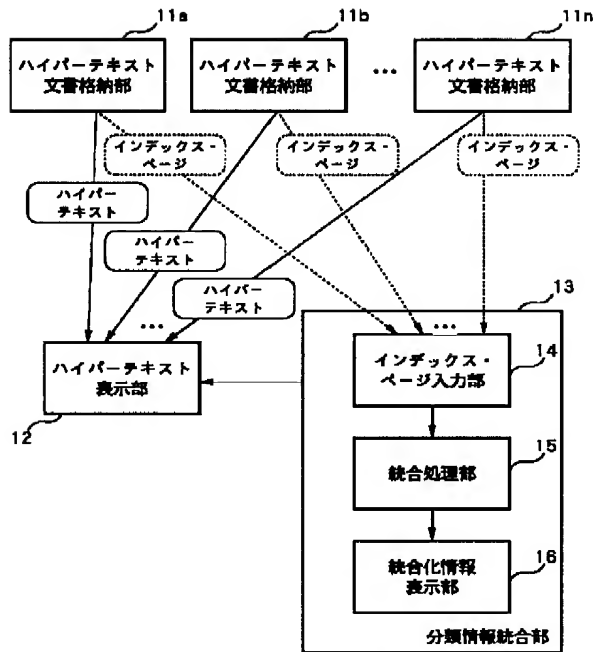
15 統合処理部

16 統合化情報表示部

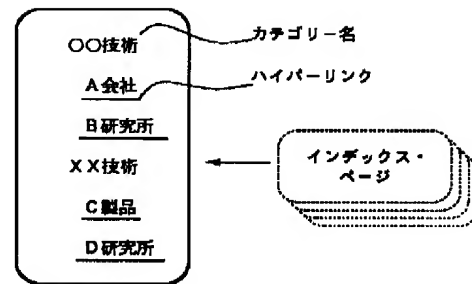
【図4】



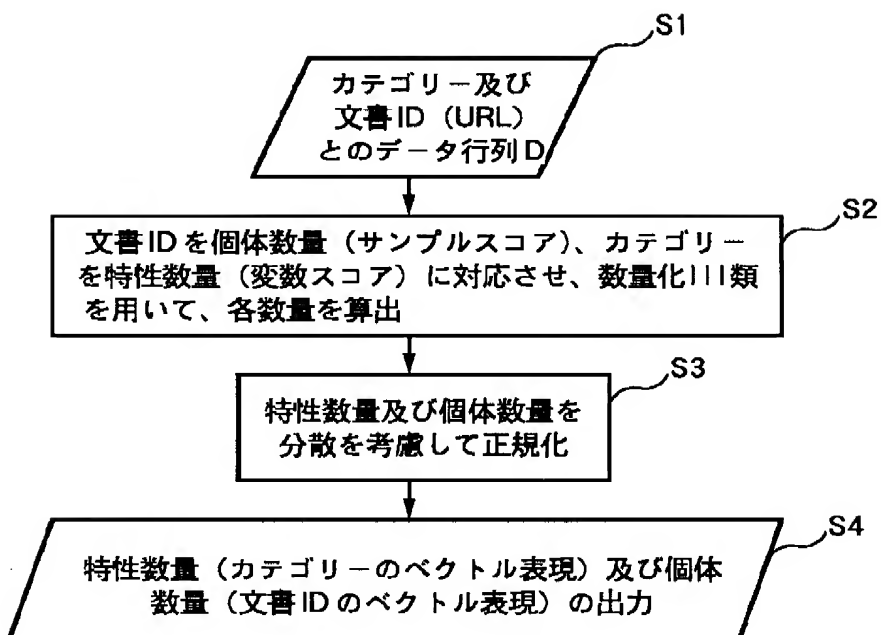
【図1】



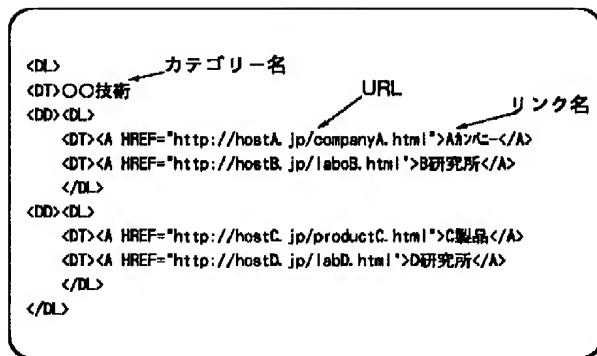
【図5】



【図2】



【図6】



【図7】

